Journal of Mathematics and Statistics 10 (1): 25-29, 2014
ISSN: 1549-3644

doi:10.3844/jmssp.2014.25.29 Published Online 10 (1) 2014 (http://www.thescipub.com/jmss.toc)

PERFORMANCE OF RIDGE REGRESSION ESTIMATOR METHODS ON SMALL SAMPLE SIZE BY VARYING CORRELATION COEFFICIENTS: A SIMULATION STUDY

[1,2,3]Anwar Fitrianto and [1]Lee Ceng Yik

[1]Department of Mathematics, Faculty of Science, Universiti Putra Malaysia, Serdang, Malaysia
[2]Institute for Mathematical Research, Universiti Putra Malaysia, Serdang, Malaysia
[3]Department of Statistics, Bogor Agricultural University, Bogor, Indonesia

Received 2013-11-05; Revised 2013-11-15; Accepted 2014-01-03

ABSTRACT

When independent variables have high linear correlation in a multiple linear regression model, we can have wrong analysis. It happens if we do the multiple linear regression analysis based on common Ordinary Least Squares (OLS) method. In this situation, we are suggested to use ridge regression estimator. We conduct some simulation study to compare the performance of ridge regression estimator and the OLS. We found that Hoerl and Kennard ridge regression estimation method has better performance than the other approaches.

Keywords: Multicollinearity, Multiple Linear Regression, Ridge Regression

1. INTRODUCTION

Main goal of multiple linear regression model is to determine the best set of parameters, β_i, so that the predicted value of dependent variables close to the real values (Orlov, 1996). In multiple linear regression models, we normally assume that the independent variables are independent. However, in practice, the explanatory variables may be correlated between each other. This inter-relation between the explanatory variables is called multicollinearity. It is the undesirable situation and can happen when the correlations among the independent variables are strong. It has several effects as has been described by Judge *et al*. (1988). One of them is that it increases the standard errors of the coefficients. In this situation, the independent assumptions are no longer valid in multiple linear regression models. The regression coefficients which are based on Ordinary Least Square estimator (OLS) tends to become unstable in the presence of multicollinearity. Wethrill (1986) also mentioned that multicollinearity is a serious problem when we make inferences for a model so that it must be handled appropriately.

Except due to strong natural linear correlation between independent variables, multicollinearity can happen due to the present of high leverage points. Recent researches focused on high leverage points correspond to outlier are Bagheri and Midi (2009).

Ridge regression estimator methods have been proposed as alternatives to the OLS estimators when the independent assumption has not been satisfied in the analysis. Several methods have been proposed for estimating the ridge parameter k and consider a criterion for comparison of the estimators. We present several methods based of ridge regression estimators.

Hence, ridge regression estimator has been proposed as an alternative to the OLS estimators when the independent assumption has not been satisfied in the analysis. The ridge estimator constrains the length of the regression coefficient of the estimator in the presence of multicollinearity. Ridge regression will be able to minimize the variance of the estimators when the design matrix is not invertible. The modification of design matrix to make its determinant different form 0 causes the estimator to be biased. This method is significantly reduces the variance of the estimators. Through this research, we want to observe how are the parameters of ridge regression estimator in different level of correlation coefficients by using Monte Carlo procedure.

Corresponding Author: Anwar Fitrianto, Department of Mathematics, Faculty of Science, Universiti Putra Malaysia, Serdang, Malaysia

1.1. Ridge Regression Model

Multicollinearity refers to a situation in which two or more predictor variables in a multiple regression model are highly correlated. Multicollinearity occurs when there is a linear relationship between one or more of the independent variables. In this situation, the regression coefficients change significantly in response to small changes in the model. The regression coefficients cannot be estimated with great accuracy because the coefficients possess large variance.

The ridge regression estimator is much more stable than the OLS estimator in the presence of multicollinearity. The ridge estimator restricts the length of the coefficients estimator in order to reduce the effects of multicollinearity (Hocking *et al*., 1976). In the presence of multicollinearity, Hoerl and Kennard (1970) introduced the ridge estimator as an alternative to the OLS estimator when the independent assumption is not longer valid. The ridge estimator is shown as follow Equation (1) (Hoerl and Kennard, 1970):

$$\hat{\beta}_k = (X'X + k\,I)^{-1} X'y \tag{1}$$

where, the I denotes an identity matrix and k is known as ridge parameter. The MSE of $\hat{\beta}_k$ is shown as follow:

$$\mathrm{MSE}\left(\hat{\beta}_k\right) = \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^2} + k^2\beta'(X'X + kI)^{-2}\beta \tag{2}$$

The $\mathrm{MSE}\left(\hat{\beta}_k\right)$ in Equation (2) depends on unknown parameters k, β and σ^2, which can't be calculated in practice. As k increase from zero to infinity, the regression estimates will approximately equal to zero. It yields minimum $\mathrm{MSE}\left(\hat{\beta}_k\right)$ compared to the OLS estimator, although these estimator results in bias, for a certain value of k (Hoerl and Kennard, 1970). In practice, we have to estimate k from the real data instead.

Standard model of a multiple linear regression can be expressed into canonical form. An orthogonal matrix D exists such that:

$$D'CD = \Lambda$$

where, C = X'X and Λ = diag ($\lambda_1, \lambda_2, \ldots, \lambda_p$) contains the eigenvalues of the matrix C, then the canonical form of the model (1) is Equation (3):

$$y = X^*\alpha + \varepsilon \tag{3}$$

where, X* = XD and α = D'β. The general form of OLS estimator is shown as follows Equation (4):

$$\hat{\alpha} = \Lambda^{-1} X^{*\prime} y \tag{4}$$

Then, the ridge estimator is written as Equation (5):

$$\hat{\alpha}(k) = \left(X^{*\prime}X^* + K\right)^{-1} X^{*\prime} y \tag{5}$$

where, K = diag ($k_1, k_2, \ldots, k_p$), $k_i > 0$. The ridge estimator in Equation (4) is known as general form of ridge regression (Hoerl and Kennard, 1970). According to Hoerl and Kennard (1970), the value of k_1 which minimizes the Equation (6):

$$\mathrm{MSE}\left(\hat{\alpha}(k)\right) = \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k_i)^2} + \sum_{i=1}^{p} \frac{{k_i}^2 {\alpha_i}^2}{(\lambda_i + k_i)^2} \tag{6}$$

Is:

$$k_i = \frac{\sigma^2}{{\alpha_i}^2} \tag{7}$$

where, σ^2 denotes the error variance of model Equation (1), α_i is the ith element of α. Equation (7) shows that the values of k_i fully depends on the unknown σ^2 and α_i. Since σ^2 and α_i are unknown, these values must be estimated from the observed data. Bhar and Gupta (2001) proposed a new criterion of detecting outlier in experimental designs which is based on average Cook-statistic. Meanwhile, Zhou and Zhu (2003) realized the fact that in practice, experiments may yield unusual observations (outliers). In the presence of outliers in a data, estimation methods such as ANOVA, truncated ANOVA, Maximum Likelihood (ML) and modified ML do not perform well, since these estimates are greatly influenced by outlier. Zhou and Zhu (2003) verified that with robust designs, one can get efficient and reliable estimates for variance components regardless of outliers which may happen in an experiment. Then Goupy (2006) conducted further research regarding outlier in an experiment who described how to discover an outlier and estimate its true value and recently, Fitrianto and Midi (2013) who compared classical and robust approach in experimental design. The method is based on the use of a dynamic variable and the ''small effects'' of the Daniel's diagram.

1.2. Method for Estimating Ridge Regression Parameters

Several methods have been proposed in order to define a new estimator that can perform better compared to the existing methods. In this part, we present some methods for estimating ridge parameter k. Hoerl and Kennard (1970) found that the best method to estimate $\hat{\alpha}(k)$ is to use $k_i = k$ for all i and they suggested k is to be $\hat{k}_{HK}$ (or HK) where Equation (8):

$$\hat{k}_{HK} = \frac{\hat{\sigma}^2}{\max(\hat{\alpha}_i)} \quad (8)$$

If σ^2 and α are known, then $\hat{k}_{HK}$ is sufficient to give ridge estimators having smaller MSE than the OLS estimator.

Hocking *et al*. (1976) defined a new method for choosing parameter k. They suggested an estimator of k by using $\hat{k}_{HSL}$ (or HSL), which produces the following estimator Equation (9):

$$\hat{k}_{HSL} = \sigma^2 \frac{\sum_{i=1}^{p} (\lambda_i \hat{\alpha}_i)^2}{\left(\sum_{i=1}^{p} \lambda_i \hat{\alpha}_i^2\right)^2} \quad (9)$$

Recently, Alkhamisi and Shukur (2007) suggested a new approach for choosing the ridge parameters k by adding $1/\lambda_{max}$ to some well-known estimators, where λ_{max} is the largest eigenvalues of X’X. They applied the modification to the previous estimator which was proposed by Hocking *et al*. (1976) in order to define a new estimator $\hat{k}_{NHSL}$ (or NHSL) Equation (10):

$$\hat{k}_{NHSL} = \sigma^2 \frac{\sum_{i=1}^{p} (\lambda_i \hat{\alpha}_i)^2}{\left(\sum_{i=1}^{p} \lambda_i \hat{\alpha}_i^2\right)^2} + \frac{1}{\lambda_{max}} = \hat{k}_{HSL} + \frac{1}{\lambda_{max}} \quad (10)$$

Since $1/\lambda_{max} > 0, \hat{k}_{NHSL}$, is greater than $\hat{k}_{HSL}$.

1.3. The use of Monte Carlo Simulation

Monte Carlo method is a stochastic technique which is used to investigate problems based on the use of random numbers and the probability statistics. We can use Monte Carlo method to solve physical problems, for example it allows us to examine more complex systems. With Monte Carlo method, we can sample the large system in a number of random configurations. Bagheri and Midi (2009) also conducted Monte Carlo simulation study in a robust approach in the presence of multicollinearity.

Simulation study will be discussed to compare the performance of ridge estimators under several degrees of multicollinearity. Different ridge estimators corresponding to different values of ridge parameter k are considered. McDonald and Galarneau (1975) and several other researchers used the following equation to generate the explanatory variables Equation (11):

$$x_{ij} = \left(1-\gamma^2\right)^{\frac{1}{2}} z_{ij} + \gamma z_{ip}, i = 1,2,...,n,\ j = 1,2,...,p \quad (11)$$

where, z_{ij} is independent standard normal pseudo-random numbers and γ is linear correlation between any two explanatory variables.

The n observations for the dependent variable y are determined by Equation (12):

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \\ &\ldots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, \ldots, n \end{aligned} \quad (12)$$

where, ε_i are independent normal $(0,\sigma^2)$ pseudo-numbers. The comparison is based on the MSE criteria. The MSEs of ridge estimators are evaluated by Equation (7).

2. METERIALS AND METHODS

2.1. Simulation Design

The simulation is conducted by using SAS release 9.2. To achieve different degrees of correlation, the explanatory variables were generated using the Equation (11). Size of sample to be considered in this research is small sample of size 20 with number of explanatory variables of equal to 10. Different values of correlations are considered in the simulation study are 0.5, 0.7 and 0.9. These three values of to represent low, moderate and high correlations between explanatory variables. The explanatory variables need to be standardized so that they will be in correlation form. Meanwhile, five different values of standard deviation to be considered in this study, which are 0.1, 0.5, 1.0, 5.0 and 10.0.

2.2. Performance Measures of the Estimators

For given values of p, σ and γ, we repeated the experiment by 1000 times. For each replication, r = 1,2,3,..,1000, the values of these three ridge estimators and the corresponding parameters, k will be estimated using the standardized variables and then the estimated coefficients are transformed back to the original model. The k values were computed based on its corresponding method.

The performance of the estimators is evaluated in terms of the averaged mean square error (Dorugade and Kashid, 2010) with the following equation:

$$\mathrm{MSE}(\hat{\alpha}_R) = \frac{1}{1000}\sum_{r=1}^{1000}\left(\hat{\alpha}_{(r)} - \alpha\right)'\left(\hat{\alpha}_{(r)} - \alpha\right)$$

The comparison between estimated MSEs are then based on the values of p, σ and γ.

3. RESULTS AND DISCUSSION

Results of the estimators performance are displayed in **Table 1**. The table displays the MSEs of each estimator under several levels of correlations corresponding to different values of σ. The first column of the table contains σ which has five different values. The second column of the table contains the correlation coefficient, γ. We compare the MSEs of each estimator under three levels of correlations where γ corresponding to different values of σ.

From **Table 1**, we noticed that the HK and HSL estimators are better than OLS estimator for all levels of correlations corresponding to different values of σ. This result in accordance and strengthens the previous research which have been conducted by Al-Hassan (2010).

Table 1. Estimated MSEs of each ridge regression estimator at three levels of correlations correspond to different values of σ

		Estimation method			
Std dev (σ)	γ	OLS	HK	HSL	NHSL
0.1	0.5	0.00004	0.0000	0.00000	0.00131
	0.7	0.00005	0.0000	0.00000	0.00021
	0.9	0.00010	0.0000	0.00000	0.00010
0.5	0.5	0.02403	0.00001	0.00004	0.01237
	0.7	0.02852	0.00004	0.00019	0.01530
	0.9	0.06486	0.00067	0.00335	0.04705
1	0.5	0.31840	0.01159	0.02860	0.15759
	0.7	0.40062	0.02624	0.07697	0.22932
	0.9	0.94697	0.15406	0.44105	0.73018
5	0.5	1.61637	0.28938	0.28713	0.85356
	0.7	2.82651	0.70882	0.81965	1.73789
	0.9	9.19491	3.64752	4.67201	7.41896
10	0.5	1.62511	0.29554	0.28489	0.8573
	0.7	2.84906	0.71761	0.80148	1.74657
	0.9	9.29982	4.02566	4.52590	7.47674

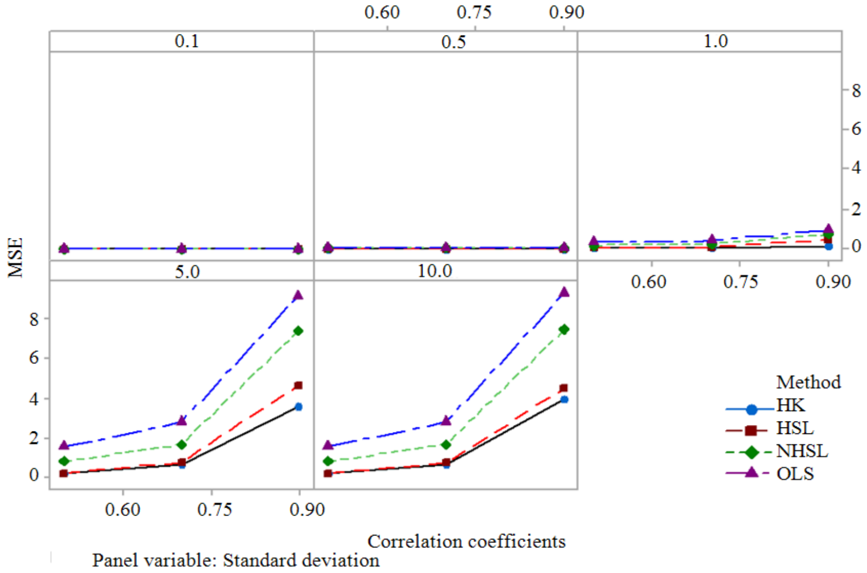

Fig. 1. Plot of estimated MSEs obtained by different ridge regression methods of each ridge regression estimator at three levels of correlations correspond to different values of σ

The MSEs of OLS estimator are lower than the MSEs of NHSL estimator for all levels of correlation when the value of σ = 0.1. However, the NHSL estimator performs better than OLS estimator when σ>0.1. From **Table 1**, we can conclude that HK estimator performs better than the OLS and other ridge estimators.

We display graphical plot of MSE (of each ridge regression method) versus the level of correlation coefficients by different values of standard deviations on **Fig. 1**. We compare the MSEs of each estimator graphically by varying the standard deviations and the correlations between the explanatory variables. The results of different levels of correlations corresponding to standard deviation as shown below.

In **Fig. 1** we can see that when the data has small variability (which are represented by σ = 0.1 and σ = 0.5) the MSE values between ridge regression method are about the same so that the small differences between them can be neglected. But when the value of standard deviation is at least one, we can observe that the MSE values increases as the standard deviation increases, regardless the methods. Moreover, we can see that the OLS estimator has the highest MSE compare to ridge estimators and within the regression estimation methods, the HK estimator performs better than the HSL and NHSL estimators for all levels of correlations.

4. CONCLUSION

In this article, we did simulation studies of several methods for estimating the ridge parameters. The performance of each ridge estimator depend on the standard deviation (σ) and the correlations between of explanatory variables (γ^2). For σ = 0.1, HK estimator and HSL estimator have smaller MSE than the OLS estimator for all levels of correlations. However, the OLS estimator is reasonably better than NHSL estimator for all levels of correlations for this given value of . HK estimator might be recommended to be used to estimate the ridge parameter k. Further investigation of ridge estimators is needed in future in order to make any definite statement.

5. ACKNOWLEDGEMENT

The reachers are grateful to University Grants Scheme by Universiti Putra Malaysia for awarding a research grant for supporting the research.

6. REFERENCES

Al-Hassan, Y.M., 2010. Performance of a new ridge regression estimator. J. Assoc. Arab Univ. Basic Applied Sci., 9: 23-26. DOI: 10.1016/j.jaubas.2010.12.006

Alkhamisi, M.A. and G. Shukur, 2007. A monte carlo study of recent ridge parameters. Commun. Stat., Simulat. Comput., 36: 535-547. DOI: 10.1080/03610910701208619

Bhar, L. and V.K. Gupta, 2001. A useful statistic for studying outliers in experimental designs. Ind. J. Stat., 63: 338-350.

Dorugade, A.V. and D.N. Kashid, 2010. Alternative method for choosing ridge parameter for regression. Applied Mathem. Sci., 4: 447-456.

Fitrianto, A. and H. Midi, 2013. A comparison between classical and robust method in a factorial design in the presence of outlier. J. Math. Stat., 9: 193-197. DOI: 10.3844/jmssp.2013.193.197

Goupy, J., 2006. Factorial experimental design: Detecting an outlier with the dynamic variable and the Daniel’s diagram. Chemomet. Intell. Laboratory Syst., 80: 156-166. DOI: 10.1016/j.chemolab.2005.05.005

Hocking, R.R., F.M. Speed and M.J. Lynn, 1976. A class of biased estimators in linear regression. Technometrics, 18: 425-437. DOI: 10.1080/00401706.1976.10489474

Hoerl, A.E. and R.W. Kennard, 1970. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12: 55-67. DOI: 10.1080/00401706.1970.10488634

Judge, G.G., R.C. Hill, W.E. Griffiths, H. Lutkepohl and T.C. Lee, 1988. Introduction to the Theory and Practice of Econometrics. 2nd Edn., John Willy and Sons, New York, ISBN-10: 0471624144, pp: 1064.

McDonald, G.C. and D.I. Galarneau, 1975. A Monte Carlo evaluation of some ridge-type estimators. J. Am. Statist. Assoc., 70: 407-412. DOI: 10.1080/01621459.1975.10479882

Orlov, M.L., 1996. Multiple Linear Regression Analysis Using Microsoft Excel. 1st Edn., Oregon State University.

Wethrill, H.H., 1986. Evaluation of ordinary ridge regression. Bull. Mathem. Statist., 18: 1-35.

Zhou, J. and H. Zhu, 2003. Robust estimation and design procedures for the random effects model. Canadian J. Stat., 31: 99-110. DOI: 10.2307/3315906

Bagheri, A. and H. Midi, 2009. Robust estimations as a remedy for multicollinearity caused by multiple high leverage points. J. Math. Statist., 5: 311-321. DOI: 10.3844/jmssp.2009.311.321